More on the value of split ballots

Fowler, Floyd Jackson Jr.

MORE ON THE VALUE OF SPLIT BALLOTS

FLOYD JACKSON FOWLER, JR.

Introduction

Much attention has been given to strategies for testing how well questions are understood and answered. This kind of evaluation has great potential for improving survey measurement. Appropriate procedures for assessing how well questions are understood and the answers are becoming increasingly common, which constitutes significant progress in survey methodology.

However, the ultimate test of whether question problems matter is how they affect the data. Although our cognitive testing strategies seem to provide meaningful information about question problems, they do not tell us how much they adversely affect data and whether or not revised questions are in fact producing "better" data.

When there are two or more candidate questions to measure a particular construct, it would be best if the answers to the candidate questions could be correlated with some kind of gold standard, a measure that there was reason to think did constitute a valid measurement of the construct itself or one to which it should be related. In the absence of that, collecting distributions of the answers to candidate questions from comparable populations, and comparing the distributions, can often provide insight into whether or not the alternative wording of questions in fact affects the data (Fowler, 2004).

The contention of this paper is that it is difficult by inspection alone to know whether or not alternative wordings of similar questions will produce different estimates and, if so, which is a "better" estimate. I present six tables that were based on such "split ballot" experiments to illustrate the point.

Methods

The data presented came from two research projects that used very similar methods. In each case, a set of questions was initially subjected to cognitive testing. Some number of volunteers from the target populations were recruited, asked test questions, then asked in

various ways to explain their understanding of the questions and how they went about answering the questions. Based on those results, when problems were found, an alternative version of the questions designed to meet the same question objectives, but with better or different question design, was developed. The data in five of the tables (Tables 1-4, 6) were derived from a small national sample of adults, identified through random digit dialing, who were randomized to one or the other versions of each test question. About 75 adults answered each version of the questions. The other example, presented in Table 5, is based on a sample of health plan members who were interviewed by telephone. In this case, about 340 respondents answered each version of the question. The analysis simply compares the distributions to the two forms of the question, looking to see if the results are the same or different depending on the question wording.

Results

Table 1 presents results for questions designed to screen adults for whether or not they had ever consumed alcoholic beverages (12 drinks in any one year) before asking a series of questions about alcohol consumption.

The first alternative asks specifically whether the respondent has ever had at least 12 drinks in any one year. Alternative 2 asks if the respondent had ever had an average of more than one drink per month.

From a mathematical standpoint, the answers should be identical. However, as the table shows, many more people said that they had 12 drinks in a year than said they had an average of one drink per month. In this case, it is not clear how accurate the responses are to Alternative 1, but it is almost certain that the answers to Alternative 2 were confounded by the notion that some respondents thought they had to have at least one drink every month in order to say "yes." That is at least part of the explanation why 18% fewer said "yes" to Alternative 2 than did so to Alternative 1.

Table 2, a follow-up question, compares two different ways of estimating how often people drink alcohol. Alternative 1 asks for the number of days in the past 30 days that the respondents had any alcoholic beverage to drink. Alternative 2 asks them to summarize in the past year how many days per week, month, or year they drank any type of alcoholic beverage. As is the case for Table 1, from a mathematical perspective, we would expect the answers to be similar. Unless the "last month" was systematically unrepresentative, which we have no reason to think was the case, the numbers reported for the last month should be similar to what the reported pattern was over the past year.

In fact, as Table 2 shows, they are not similar at all. The average over the past year results in many more days of reported drinking than the question about the past month. Part of the reason may be that those people who do drink sometimes, but not often, did not report that they did not drink at all during the year but might report no days in the past month. However, if that was a main factor, the bottom two categories should add up to a similar sum, and they do not.

Table 3 is an examination of the effect of giving examples when abstract terms are used in questions. In this case, the question is about days respondents did any strenuous activities in and around their home. The original question provides a number of examples of strenuous activities, while the alternative question leaves out the examples. In all other respects, the questions are the same. As can be seen from the data, providing the examples greatly increases the number of days on which the respondents report any strenuous activity.

The data in Table 4 provide a comparison that might seem similar to Table 3. In this case, the question is about dental care. The original question provides respondents with examples of the various kinds of dentists they might have seen, such as orthodontists or oral surgeons. In contrast, the alternative assumes that respondents know what dental care means and provides no further examples. In contrast to the results in Table 3, in Table 4 there is no difference in the rates at which visits for dental care are reported. Providing the examples of kinds of dentists and dental care has essentially no effect on the answers.

Table 5 shows two series of questions designed to identify people who have chronic health conditions. In this case, for both series of questions, a chronic health condition was one that had lasted for at least three months and that either had required the taking of prescription medicines for three months or had led to seeing a doctor about the condition three or more times in the past year. The series differ in one crucial respect. The "standard" form begins with a question about whether or not the respondent has had a condition for three months, then asks if any of those conditions met the standard for either use of prescription drugs or seeking medical care. The alternative begins with whether or not respondents have had a condition for which they have taken medication for three months or seen a doctor three or more times in the past year, then asks whether the condition has lasted for at least three months. As the table shows, the series that begins with the behavioral implications of the condition, taking drugs or seeing a doctor, produce reports of many more chronic conditions than the series that started with the general question about having a condition for three or more months. In this case, we have cognitive testing results that show that people are not consistent in their understanding of what a "physical or medical condition" is. It seems highly likely that there is significant

under-reporting of conditions based on the ambiguity of the question. Hence, while we are not certain about the level of validity of the alternative, we are pretty sure, based on our testing, that the answers to the alternative are better than those to the "standard" series.

Table 6 shows two series of questions designed to identify people who had been injured in an automobile accident "because of their driving." The alternative series breaks the initial question into three parts: 1) injured in an automobile accident, 2) while you were driving, 3) because of your driving. The data show that the estimates from the two approaches are significantly different. Many fewer people end up saying "yes" when they are asked the three-question series than when they are asked the question in its initial form. Again, because of initial testing of this question, we are pretty sure that respondents did not attend to all of the issues raised when they are all presented in a single question. Based on that analysis, we are confident that the second series of questions is producing more valid data.

Discussion

The take-away point from the above is that split ballot comparisons of alternative forms of questions provide invaluable information about how question wording affects resulting data. For some of the above examples, such as Tables 5 & 6, we had a strong theory based on cognitive testing about why one of the versions might contain significant error. However, we needed the split-ballot data to prove it.

Tables 3 & 4 present similar comparisons but with different results. In Table 3, providing examples of strenuous activities greatly increased the reporting of such activities; in contrast, in Table 4, giving examples of dental care had no effect on the amount of dental care that was report. I would argue that it would be very difficult for even a skilled question design expert to have reliably predicted in advance how those two comparisons would turn out.

Tables 1 & 2, comparing alternative ways of asking about alcohol consumption, provide great examples of hard-to-predict results. Logically and mathematically, the data should be identical from those two pairs of questions. In fact, the results are very different.

Thus, cognitive testing of questions provides an extremely useful way to identify question problems, to diagnose problems with questions, and stimulate the revision of survey questions. However, in the end, we need to have data about how the resulting survey estimates will be affected by problems in our original or revised questions. The point of this paper is to encourage researchers to build in split-ballot tests of their proposed

question wording prior to fielding their full-scale surveys. Cognitive testing results in combination with split-ballot testing provides a better basis for making decisions about which questions to ask.

Acknowledgements

The collection of the data represented in this paper was funded by contract from the National Center for Health Statistics (NCHS) and by a cooperative agreement between the Agency for Healthcare Research and Quality and the Harvard Medical School (Grant # HS-09205). The author wants acknowledge to contribution of Paul Beatty at NCHS to the design of some of these experiments. In addition, it should be noted that some of these results were reported at the Questionnaire Development Evaluation and Testing (QDET) conference and will be published in the monograph derived from the conference (Fowler, 2004).

References

Fowler, F.J. (2004) "Getting Beyond Pretesting and Cognitive Interviews: The Case for More Experimental Pilot Studies" In Presser, S. et. al. (eds) *Questionnaire Development Evaluation and Testing Methods*, New York: Wiley.

Table 1:

Alternative 1

The next questions are about drinking alcoholic beverages. Included are liquor, such as whiskey or gin, beer, wine, wine coolers, and any other type of alcoholic beverage. In any one year, have you ever had at least 12 drinks of any type of alcoholic beverage?

Alternative 2

The next questions are about drinking alcoholic beverages. Included are liquor, such as whiskey or gin, beer, wine, wine coolers, and any other type of alcoholic beverage. In any one year, have you ever had an average of more than one drink per month?

	Alternative 1	Alternative 2
Yes	71%	53%
No	29%	47%
Total	100% (n=77)	100% (n=79)

$p < .02$

Table 2:

Alternative 1

In the last 30 days, on how many days did you drink any type of alcoholic beverage?

Alternative 2

In the past year, on how many days per week, month, or year did you drink any type of alcoholic beverage?

Number of Days	Alternative 1	Alternative 2
0	46%	25%
1-5	41%	45%
6-10	7%	17%
11+	6%	13%
Total	100% (n=79)	100% (n=75)
Mean Days	2.6	5.3

$p < .01$

Table 3:

Original

During the past 30 days, on how many days did you do strenuous tasks in or around your home? By strenuous tasks, we mean things such as shoveling soil in a garden, chopping wood, major carpentry projects, cleaning the garage, scrubbing floors, or moving furniture.

Alternative

During the past 30 days, on how many days did you do any strenuous tasks in or around your home?

Number of Days	Original	Alternative
0	32%	42%
1-5	34%	37%
6-10	13%	10%
11+	21%	11%
Total	100% (n=77)	100% (n=79)
Mean Days	4.66	2.72

$p < .05$

Table 4:

Original

About how many months has it been since you last saw or talked to a dentist? Include all types of dentists, such as orthodontists, oral surgeons, or all other dental specialists, as well as dental hygienists.

Alternative

About how many months has it been since you last went to a dentist office for any type of dental care?

	Original	Alternative
6 months or less	60%	57%
More than 6 months but not more than 1 year	14%	18%
More than 1 year	26%	25%
Total	100% (n=77)	100% (n=79)

NS

Table 5:

Standard

a. Do you now have any physical or medical conditions that have lasted *for at least 3 months*? (Women: DO NOT include pregnancy.)

b. In the last 12 months, have you *seen a doctor* or other health provider *more than twice* for any of these conditions?

c. Have you been taking prescription medicine for at least 3 months for any of these conditions?

Alternative

a. In the past 12 months, have you seen a doctor or other health provider 3 or more times for the same condition or problem?

b. Is this a condition that has lasted for at least 3 months? (Do *not* include pregnancy.)

c. Do you now need to take medicine prescribed by a doctor (other than birth control)?

d. Is this to treat a condition that has *lasted for at least 3 months*? (Do not include pregnancy or menopause.)

	Has Chronic Condition	
Yes	38%	56%
No	62%	44%
Total	100% (335)	100% (347)

$p < .01$

Table 6:

Original

This question is about automobile injuries, including injuries from crashes, burns, and any other kind of accidents. Have you ever had an injury because of your driving?

Alternative

a. This question is about automobile injuries, including injuries from crashes, burns, and any other kind of accidents. Have you ever had an injury while you were in a car?

b. Were you ever the driver when you were injured?

c. Were you ever injured because of your driving?

	Original	**Alternative**
Yes	8%	2%
No	92%	98%
Total	100% (n=79)	100% (n=77)

$p < .05$

Contact

Floyd Jackson Fowler, Jr.
Center for Survey Research
University of Massachusetts, Boston
100 Morrissey Blvd.
Boston, MA 02125
U.S.A.
email: floyd.fowler@umb.edu